5_6_Magic_BLAST_contig_subsets_filtered

July 30, 2021

[1]:
```
import numpy as np
import math
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from matplotlib.ticker import MaxNLocator
from pandas.plotting import scatter_matrix
import pathlib
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:95% !important; }</style>"))
```

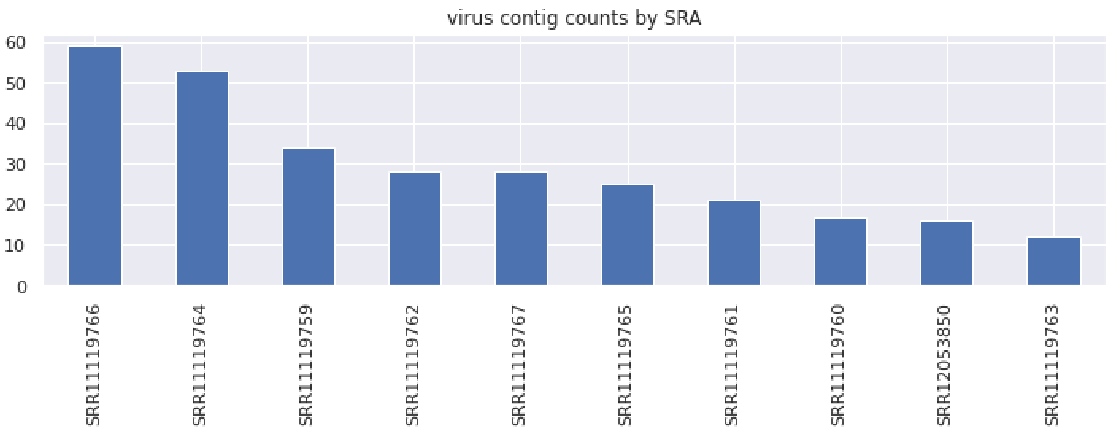
```
<IPython.core.display.HTML object>
```

[3]:
```
PROJECT_CODE='PRJNA607174'
BASE_PATH = f'/mnt/1TB_0/Data/Assembly/{PROJECT_CODE}/'

dbname='nt'
kmer='k141'

f_contigs_file_tail=f'_{dbname}_magic_blast_asc_contigs.txt'
```

[4]:
```
subsets=['vector','virus']
```

[71]:
```
path = BASE_PATH+'/contig_subsets/virus/'
df=pd.read_csv(path+'PRJNA607174_contig_BLAST_analysis_virus_magicblast_nt.
 ↪tsv', sep='\t')
```

[83]:
```
df['sra'].value_counts().plot(kind='bar', figsize=(10,4))
plt.title('virus' +' contig counts by SRA')
plt.tight_layout()
plt.savefig(path+'PRJNA607174_contig_BLAST_analysis_virus_magicblast_nt_sras.
 ↪png')
plt.show()
```



[84]:

```
asc=df['accession']
```

[85]:

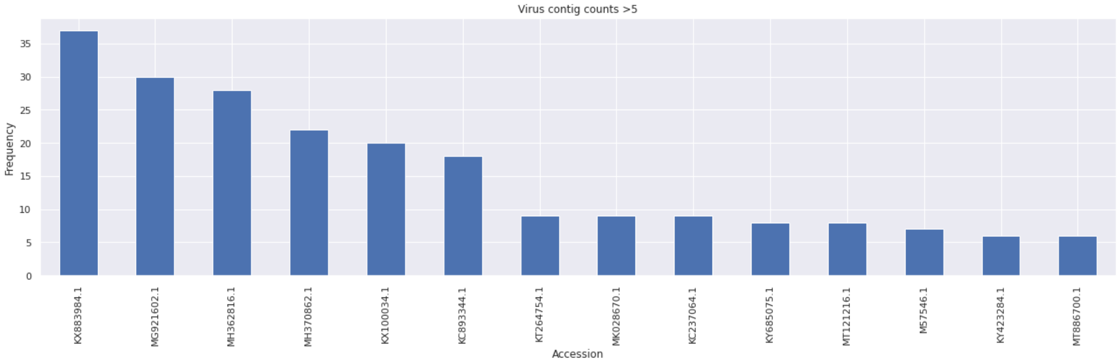
```
asc.unique()
```

[85]:

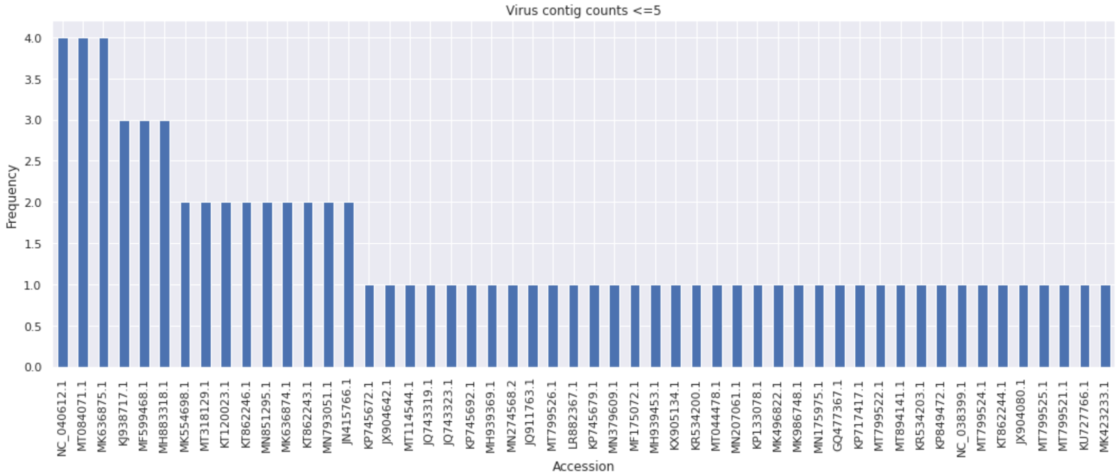
```
array(['GQ477367.1', 'JN415766.1', 'JQ743319.1', 'JQ743323.1',
       'JQ911763.1', 'JX904080.1', 'JX904642.1', 'KC237064.1',
       'KC893344.1', 'KJ938717.1', 'KP133078.1', 'KP717417.1',
       'KP745672.1', 'KP745679.1', 'KP745692.1', 'KP849472.1',
       'KR534200.1', 'KR534203.1', 'KT120023.1', 'KT264754.1',
       'KT862243.1', 'KT862244.1', 'KT862246.1', 'KU727766.1',
       'KX100034.1', 'KX883984.1', 'KX905134.1', 'KY423284.1',
       'KY685075.1', 'LR882367.1', 'M57546.1', 'MF175072.1', 'MF599468.1',
       'MG921602.1', 'MH362816.1', 'MH370862.1', 'MH883318.1',
       'MH939369.1', 'MH939453.1', 'MK028670.1', 'MK423233.1',
       'MK496822.1', 'MK554698.1', 'MK636874.1', 'MK636875.1',
       'MK986748.1', 'MN175975.1', 'MN207061.1', 'MN274568.2',
       'MN379609.1', 'MN793051.1', 'MN851295.1', 'MT044478.1',
       'MT084071.1', 'MT114544.1', 'MT121216.1', 'MT318129.1',
       'MT799521.1', 'MT799522.1', 'MT799524.1', 'MT799525.1',
       'MT799526.1', 'MT886700.1', 'MT894141.1', 'NC_038399.1',
       'NC_040612.1'], dtype=object)
```

[86]:

```
ax = df['accession'].value_counts().loc[lambda x : x>5].plot(kind='bar',
                                    figsize=(18,6),
                                    title="Virus contig counts >5")
ax.set_xlabel("Accession")
ax.set_ylabel("Frequency")
plt.tight_layout()
plt.
 ↪savefig(path+'PRJNA607174_contig_BLAST_analysis_virus_magicblast_nt_accessions_>5.
 ↪png')
```



[82]:
```
ax = df['accession'].value_counts().loc[lambda x : x<=5].plot(kind='bar',
                                                    figsize=(18,6),
                                                    title="Virus contig counts <=5")
ax.set_xlabel("Accession")
ax.set_ylabel("Frequency")
plt.
 ↪savefig(path+'PRJNA607174_contig_BLAST_analysis_virus_magicblast_nt_accessions_<=5.
 ↪png')
```



0.1 vectors

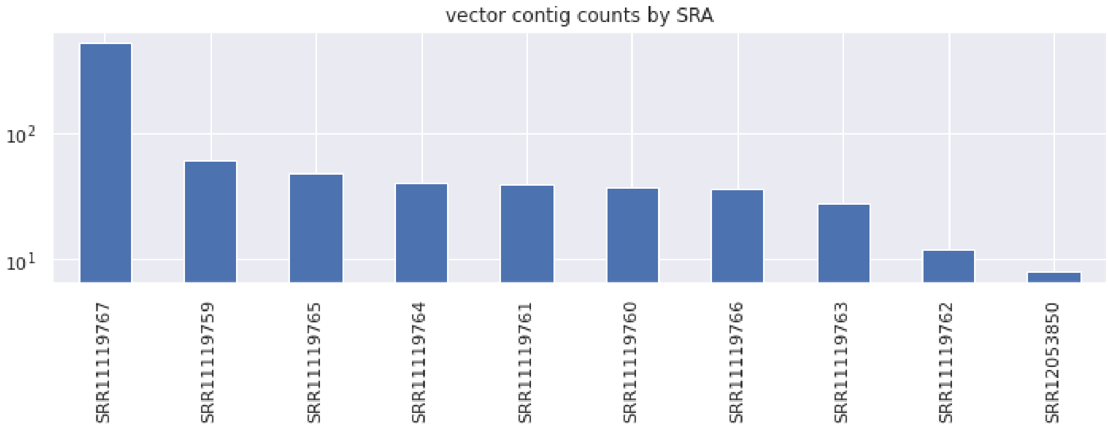
[96]:
```
path = BASE_PATH+'/contig_subsets/vector/'
df=pd.read_csv(path+'PRJNA607174_contig_BLAST_analysis_vector_magicblast_nt.
 ↪tsv', sep='\t')
df.head()
```

[96]:
```
           sra       contig   accession  \
0  SRR11119760      k141_914  AB545980.1
1  SRR11119761  k141_1166153  AB570081.1
2  SRR11119767    k141_873975  AB609713.1
3  SRR11119767    k141_873975  AB609714.1
4  SRR11119767    k141_873975  AB609715.1

                                         description     cigar
0  Cloning vector pTip-istAB-sacB DNA, complete s…      719M
1  Cloning vector pInSRT-GFPV1 DNA, complete sequ…      370M
2  Gene trap vector pCMT-SAhygpA-NP21 DNA, comple…  105S246M
3  Gene trap vector pCMT-SAhygpA-NP22 DNA, comple…  105S246M
4  Tol2 transposon-based gene trap vector pT2F2-S…  105S246M
```

[98]:
```python
descr=df['description']
descr.unique()
```

[98]:
```
array(['Cloning vector pTip-istAB-sacB DNA, complete sequence',
       'Cloning vector pInSRT-GFPV1 DNA, complete sequence',
       'Gene trap vector pCMT-SAhygpA-NP21 DNA, complete sequence',
       'Gene trap vector pCMT-SAhygpA-NP22 DNA, complete sequence',
       'Tol2 transposon-based gene trap vector pT2F2-SAhygpA-NP21 DNA, complete
sequence',
       'Tol2 transposon-based gene trap vector pT2F2-SAhygpA-NP22 DNA, complete
sequence',
       'Escherichia coli-Bacteroides shuttle vector pVAL-1 DNA, complete
sequence',
       'Escherichia coli-Bacteroides shuttle vector pTIO-1 DNA, complete
sequence',
       'Homo sapiens 12 NOVECTOR RP11-1006M13 () complete sequence',
       'Homo sapiens 3 NOVECTOR RP11-784F16 (Roswell Park Cancer Institute Human
BAC Library) complete sequence',
       'Homo sapiens 12 NOVECTOR RP11-627D10 (Roswell Park Cancer Institute
Human BAC Library) complete sequence',
       "Bos taurus NOVECTOR CH240-248M14 (Children's Hospital Oakland Research
Institute Bovine BAC Library complete sequence",
       'Shuttle cosmid vector pFD666, complete sequence',
       'DNA-binding vector pODB80, complete sequence',
       'Eukaryotic expression vector pCR3.1mBCL-XL, complete sequence',
       'Cloning vector pERV3, complete sequence',
       'Expression vector pARA13, complete sequence',
       'Integration vector pCD11PZ1 chloramphenicol transacetylase (cat) and
beta-galactosidase (lacZ) genes, complete cds',
       'Cloning vector pdeltaE1sp1A(CMV-GFP), complete sequence',
       'Cloning vector pJC1', "Cloning vector pCAT-3', complete sequence",
       'Monster GFP vector phMGFP, complete sequence',
       'Cloning vector pMK2016, complete sequence',
       'Cloning vector pTCCR, complete sequence',
       'Cloning vector pTCCR-Auto, complete sequence',
       'Expression vector pBeloBacModified, complete sequence',
       'Cloning vector pOriR6K-zeo-ie, complete sequence',
       'Expression vector pHT2, complete sequence',
       'Saccharomyces cerevisiae expression vector p426GPD, complete sequence',
       'Saccharomyces cerevisiae expression vector p423GALL, complete sequence',
       'Shuttle vector pPW380, complete sequence',
       'Inducible shuttle vector pPW578, complete sequence',
       'Microarray spiking control vector pSP64bb1sp1, complete sequence',
       'Microarray spiking control vector pSP64bb2sp3, complete sequence',
       'Microarray spiking control vector pSP64bb2sp4, complete sequence',
       'Microarray spiking control vector pSP64bb2sp5, complete sequence',
       'Cloning vector pO6T, complete sequence',
       'Cloning vector pOri1, complete sequence',
       'Cloning vector pCW, complete sequence',
       'Cloning vector pCW1, complete sequence',
       'Cloning vector pCW2, complete sequence',
       'YAC construction vector pRML1, complete sequence',
       'Expression vector pET32a-LIC, complete sequence',
       'Cloning vector pEFBOS-IRESGFPNeo, complete sequence',
       'Cloning vector pJM2, complete sequence',
       'Expression vector pHsh-kan, complete sequence',
       'Cloning vector pPAC7, complete sequence',
       'Cloning vector for bidirectional expression pPMV ampR gene for
ampicillin resistance protein',
       'Shuttle vector pSenLys', 'Shuttle vector pSenArg',
       'Shuttle vector pSenSer', 'Shuttle vector pSenOAS',
       'Acinetobacter baumannii expression vector pAT-RA, complete sequence',
       'Cloning vector pori6K-pA, complete sequence',
       'Cloning vector pUC57-Kan, complete sequence',
       'Cloning vector pBG1805, complete sequence',
       'Cloning vector pRAB11, complete sequence',
       'Cloning vector pNHG, complete sequence',
       'Shuttle vector pLV.CAG.hrGFP, complete sequence',
       'Cloning vector pNHG-CapNM, complete sequence',
       'Cloning vector pLV.donor.eGFP, complete sequence',
       'BAC cloning vector pDEV-vac, complete sequence',
       'Expression vector pMVAX1(c), complete sequence',
       'Cloning vector pTT-PB-SOKMLNpuro, complete sequence',
       'Cloning vector pCMV-TALER31, complete sequence',
       'Cloning vector BASIC_13_ATG-GFP-nostop, complete sequence',
       'Cloning vector BASIC_3_Kan, complete sequence',
       'Cloning vector BASIC_4_Cm, complete sequence',
       'Cloning vector BASIC_5_Amp, complete sequence',
       'Cloning vector BASIC_6_J23102, complete sequence',
       'Cloning vector BASIC_8_RBS34-mcherry, complete sequence',
        'Cloning vector BASIC_9_ATG-mcherry, complete sequence',
        'Cloning vector BASIC_10_J23101-RBS32-GFP, complete sequence',
        'Cloning vector BASIC_11_RBS34-GFP, complete sequence',
        'Cloning vector BASIC_12_ATG-GFP, complete sequence',
        'Cloning vector BASIC_7_J23101-RBS34-mcherry, complete sequence',
        'Cloning vector pVK9PtacMuAB, complete sequence',
        'Expression vector pOPT-GST_tjp1aC, complete sequence',
        'Expression vector pOPT-GST_tjp1aN, complete sequence',
        'Clostridium acetobutylicum gene inactivation vector pHKO1, complete
sequence',
        'Cloning vector pBAC-LC, complete sequence',
        'Cloning vector pBAC-TC, complete sequence',
        'Cloning vector pTT-PB-SOKM-puro, complete sequence',
        'Cloning vector pTT-PB-hTERT-puro, complete sequence',
        'Expression shuttle vector pGJ103, complete sequence',
        'Cloning vector PRFP, complete sequence',
        'Cloning vector pGZ-DSB-CO, complete sequence',
        'Expression vector pcDNA-NLS-PhoCl-mCherry, complete sequence',
        'Expression vector pcDNA-PhoCl-mCherry-myc, complete sequence',
        'Cloning vector pLCP.CFH, complete sequence',
        'Cloning vector pNR.CFH, complete sequence',
        'Expression vector pInSRT-hM(V1) DNA, complete sequence',
        'Expression vector pUC57-Amp-aac(3)-Ia DNA, complete seuence',
        "Expression vector pUC57-Amp-aac6'Ibcr DNA, complete seuence",
        "Expression vector pUC57-Amp-aac2'-Ia DNA, complete seuence",
        "Expression vector pUC57-Amp-ant2''Ia DNA, complete seuence",
        "Expression vector pUC57-Amp-aph2''Ia DNA, complete seuence",
        "Expression vector pUC57-Amp-aph3'Ia DNA, complete seuence",
        'Expression vector pUC57-Amp-armA DNA, complete seuence',
        'Expression vector pUC57-Amp-rmtA DNA, complete seuence',
        'Expression vector pUC57-Amp-npmA DNA, complete seuence',
        'Mammalian expression vector pDsRed-N-GW, complete sequence',
        'Mammalian vector pCFP-N-GW, complete sequence',
        'Vector pGFP-N-GW, complete sequence',
        'Vector pGST-N-GW, complete sequence',
        'Mammalian vector pCR3-GW-GST-C, complete sequence',
        'Mammalian vector pCR3-EGFPC-GW, complete sequence',
        'Mammalian vector pCR3-ECFPC-GW, complete sequence',
        'Mammalian vector pCR3-EYFPC-GW, complete sequence',
        'Vector pSP72-E-hA20, complete sequence',
        'Mammalian expression vector pCR3-pro-IL-1B-D117A, complete sequence',
        'Mammalian expression vector pCR3-pro-IL-1B-D108A, complete sequence',
        'Mammalian expression vector pCR3-pro-IL-1B-D105A, complete sequence',
        'Mammalian expression vector pCR3-pro-IL-1B, complete sequence',
        'Yeast expression vector pGBKT7-Cezanne(17-858), complete sequence',
        'Yeast expression vector pGBKT7-Cezanne(17-443), complete sequence',
        'Yeast expression vector pGBKT7-Cezanne(444-858), complete sequence',
        'Mammalian expression vector pcDNA3-HA-IkappaBalpha-superrepressor,
 complete sequence',
        'Mammalian expression vector pCR3-FLAGhRIP1, complete sequence',
        'Shuttle vector pMSK1-E, complete sequence',
        'Shuttle vector pMSK1-E-D565A, complete sequence',
        'Mammalian expression vector pEGFPC2-haEctn(69-3454), complete sequence',
        'Mammalian expression vector pCMF2E-hFADD-DED, complete sequence',
        'Mammalian expression vector pCMF2E-hFADD, complete sequence',
        'Mammalian expression vector pCMF2E-hFADD-DN, complete sequence',
        'Mammalian expression vector pCMF2E-hRAIDD, complete sequence',
        'Mammalian expression vector pCMF2E-mCASP-2, complete sequence',
        'Mammalian expression vector pCMF2E-mCASP-8, complete sequence',
        'Mammalian expression vector pCMF2E-link-mCASP-8, complete sequence',
        'Mammalian expression vector pCMF2E-mCASP-2-C320A, complete sequence',
        'Mammalian expression vector pcDNA1-hFADD, complete sequence',
        'Mammalian expression vector pcDNA1-hFADD-DED, complete sequence',
        'Mammalian expression vector pcDNA1-hFADD-V121N, complete sequence',
        'Mammalian expression vector pCMF2E-link-hFADD, complete sequence',
        'Mammalian expression vector pCMF2E-link-hFADD-DN, complete sequence',
        'Mammalian expression vector pCMF2E-link-hFADD-DED, complete sequence',
        'Mammalian expression vector pcDNA1-hTRADD, complete sequence',
        'Mammalian expression vector pcDNA1-hFADD-DN, complete sequence',
        'Mammalian expression vector pcDNA1-hTNFR55-E, complete sequence',
        'Mammalian expression vector pCDM9, complete sequence',
        'Mammalian expression vector pCDM8-His-PKR, complete sequence',
        'Mammalian expression vector pCDM8-PKR, complete sequence',
        'Mammalian expression vector pEGFPC2-haNctn(39-2885), complete sequence',
        'Mammalian expression vector pEGFPC2-haTctn(179-2860), complete
 sequence',
        'Mammalian expression vector pcDNAhRAIDD-E, complete sequence',
        'Mammalian expression vector pCDM92, complete sequence',
        'Yeast expression vector pGAL1PNiST-1, complete sequence',
        'Mammalian expression vector pCR3VSV-CED-4, complete sequence',
        'Mammalian expression vector pCR3FLAG-CED-3, complete sequence',
        'Mammalian expression vector pCR3Flag-E8-FLIP, complete sequence',
        'Mammalian expression vector pCDM8hFasR, complete sequence',
        'Mammalian expression vector pCDM8, complete sequence',
        'Yeast expression vector pSCGAL10-SN, complete sequence',
        'Vector pEBSV1d, complete sequence',
        'Vector pEBSV3d, complete sequence',
        'Mammalian expression vector pBM6DraA6, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MPEG1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHB11, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHA12, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHB7, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGB4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGB1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PLEKHG2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-TSHZ3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGC4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGC3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-DISP1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF217, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-GRIN2A, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-DUSP21, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MEAF6, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-NANOS3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-HIST1H1C, complete
sequence',
        'Human ORFeome Gateway entry vector pENTR223-P2RY11, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-HIST1H1A, complete
sequence',
        'Human ORFeome Gateway entry vector pENTR223-IER2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-NANP, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-C5orf20, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ALPK1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-HOXB13, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-RPS14P3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-TAS2R60, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-APOF, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF488, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-INHBE, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-FFAR1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR1D2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-SPRY4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-C1orf135, complete
sequence',
        'Human ORFeome Gateway entry vector pENTR223-ARPM1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAS1L, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-CCBP2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-FAM46C, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-DDI1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-SGPP2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ELK3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-REXO4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-EVI2B, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAB21L2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-SPZ1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF597, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-FAM46B, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PARS2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZCCHC12, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAP6, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-NYX, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OLFM4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF280A, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-EGR1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-DNHD1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-SLC22A2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-AHNAK2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-AMY2A, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-HAS2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ARSI, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-C5orf54, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-GPR37, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-KCNA5, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAP3K13, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-KIF2B, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDH20, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-IRS1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-LINC00115, complete
 sequence',
        'Human ORFeome Gateway entry vector pENTR223-TP73-AS1, complete
 sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR6W1P, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-C15orf59, complete
 sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR10K1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR56B4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR51E1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-LRRC52, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR51F2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR52N4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-REP15, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-OR13C3, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-LOC407835, complete
 sequence',
        'Human ORFeome Gateway entry vector pENTR223-HYAL4, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-BHLHA15, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-FZD10, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF574, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-NHSL2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZBTB39, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGA6, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-LRFN2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PCDHGB5, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAGEE1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-MAML2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-RANBP6, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-TCHHL1, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-ZNF804B, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-PNLIP, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-DCAF12L1, complete
 sequence',
        'Human ORFeome Gateway entry vector pENTR223-RNF25, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-URB2, complete sequence',
        'Human ORFeome Gateway entry vector pENTR223-SLX4, complete sequence',
        'Cloning vector pBGS9+ with the kanamycin resistance gene from Tn903,
 lacZ alpha-peptide, pBR322 origin, and the SspI fragment of bacteriophage f1',
        'Cloning vector pBGS9- with the kanamycin resistance gene from Tn903,
 lacZ alpha-peptide of pUC9 and SspI fragment of bacteriophage f1',
        'Cloning vector pVWEx1, complete sequence',
        'Cloning vector pVK-lacIQ-Ptac-MuAB, complete sequence',
        'Shuttle vector PCY, complete sequence',
        'Cloning vector pTLD67, complete sequence',
        'Cloning vector pTLD39, complete sequence',
        'Expression vector dCas9g1pac, complete sequence',
        'Gateway cloning destination vector pcGFP1Fpac, complete sequence',
        'Lentiviral vector pTK1940, complete sequence',
        'Vector pBud.tTA.mCherry, complete sequence',
        'Protoplast Transient Overexpression Vector pRTVnYFP terminator and
 promoter region',
        'Protoplast Transient Overexpression Vector pRTVcGFP, complete sequence',
        'Plant Stable Overexpression Vector pRHVcGFP, complete sequence',
        'Plant Stable Overexpression Vector pRGVcGFP, complete sequence',
        'Cloning vector pBac[AttB-3xP3::RFP-dsx_homology-AttP-3xP3::GFP-AttP-
 dsx_homology-AttB], complete sequence',
        'Cloning vector pSGKP-km, complete sequence',
        'Cloning vector pSGKP-spe, complete sequence',
        'Expression vector pYL4, complete sequence',
        'Vector LeuRS.Ad, complete sequence',
        'Cloning vector pRG_Duet1, complete sequence',
        'Cloning vector CRISPRi pRG_dCas9Pminus, complete sequence',
        'Cloning vector CRISPRi pRG_dCas9, complete sequence',
        'Cloning vector CMV-DHFR-mCherry, complete sequence',
        'Cloning vector CMV-mWTGFP, complete sequence',
        'Cloning vector CMV-DHFR-mGFP, complete sequence',
        'Cloning vector pVAX1-BMP2, complete sequence',
        'Cloning vector pVAX1-BMP7, complete sequence',
        'Cloning vector pVAX1-BMP2/7 (-), complete sequence',
        'Cloning vector pVAX1-BMP2/7 (+), complete sequence',
        'Cloning vector pET28A-blaBIC-1, complete sequence',
        'Cloning vector RaV_infectious_clone_(pT7-RaV), complete sequence',
        'Cloning vector pUD1074, complete sequence',
        'Expression vector pVG185_w2-y1, complete sequence',
        'Expression vector pVG307_truncated, complete sequence',
        'Expression vector pVMG47_truncated-L, complete sequence',
        'Expression vector pVMG48_truncated-R, complete sequence',
        'Cloning vector AGG1533, complete sequence',
        'Cloning vector pCasCure-Apr, complete sequence',
        'Cloning vector pCasCure-Rif, complete sequence',
        'Yeast display vector pNMB2, complete sequence',
        'Yeast display vector pNMB1m_C_Nb207, complete sequence',
        'Yeast display vector pNS1MB, complete sequence',
        'Cloning vector pVG344_Aste_kh2-MCRv3-Vasa-Cas9_SWAP, complete sequence',
        'Recombinant vector AAVCOVID19-1, complete sequence',
        'Recombinant vector AAVCOVID19-3, complete sequence',
        'Yeast CUP1 expression/integration cloning vector YITAG100 with the
 hemagglutinin tag sequence, complete sequence',
        'Yeast CUP1 expression-multicopy (2micron) cloning vector YRTAG300 with
 the hemagglutinin tag sequence, complete sequence. selection)',
        'Phagemid vector pBK-CMV, complete sequence',
        'Cloning vector pODB8, GAL4 DNA-binding domain vector, complete
 sequence',
        'Plasmid pMK20 cloning vector', 'pWE15 cosmid vector DNA',
        'Cloning vector pSP64 (polyA)', 'pWE15A cosmid vector DNA'],
       dtype=object)
```

[89]:
```
df['sra'].value_counts().plot(kind='bar', figsize=(10,4))
plt.title('vector' +' contig counts by SRA')
plt.yscale('log')
plt.tight_layout()
plt.savefig(path+'PRJNA607174_contig_BLAST_analysis_vector_magicblast_nt_sras.
 ↪png')
plt.show()
```



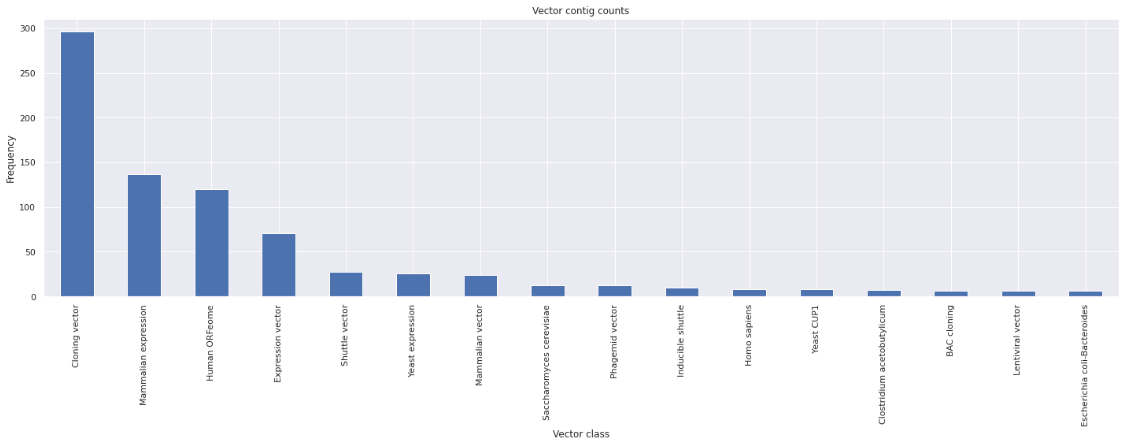
[90]:
```
def proc(s):
    l = s.split()
    return ' '.join(l[:2])
```

[91]:
```
df['class'] = [proc(s) for s in df['description'].values.tolist()]
```

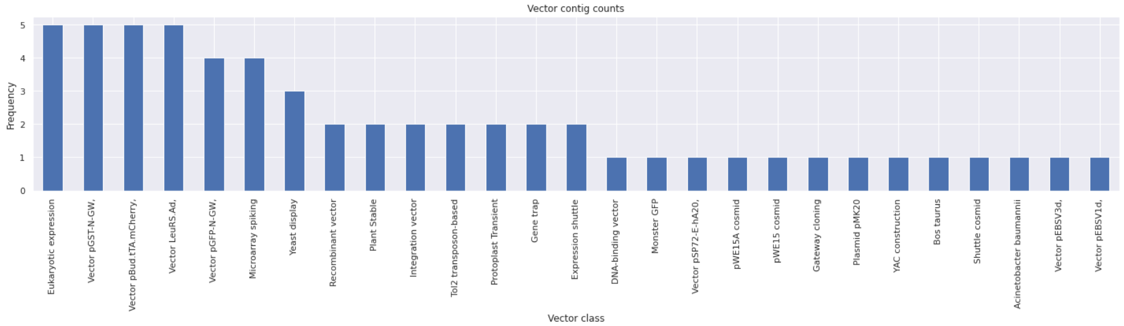
[92]:
```
df.head()
```

[92]:
```
           sra         contig    accession  \
0  SRR11119760      k141_914  AB545980.1
1  SRR11119761  k141_1166153  AB570081.1
2  SRR11119767   k141_873975  AB609713.1
3  SRR11119767   k141_873975  AB609714.1
4  SRR11119767   k141_873975  AB609715.1

                                         description     cigar  \
0  Cloning vector pTip-istAB-sacB DNA, complete s…      719M
1  Cloning vector pInSRT-GFPV1 DNA, complete sequ…      370M
2  Gene trap vector pCMT-SAhygpA-NP21 DNA, comple…  105S246M
3  Gene trap vector pCMT-SAhygpA-NP22 DNA, comple…  105S246M
4  Tol2 transposon-based gene trap vector pT2F2-S…  105S246M

                  class
0        Cloning vector
1        Cloning vector
2             Gene trap
3             Gene trap
4  Tol2 transposon-based
```

[95]:
```
ax = df['class'].value_counts().loc[lambda x : x>5].plot(kind='bar',
                                    figsize=(20,8),
                                    title="Vector contig counts")
ax.set_xlabel("Vector class")
ax.set_ylabel("Frequency")
plt.tight_layout()
plt.
 ↪savefig(path+'PRJNA607174_contig_BLAST_analysis_vector_magicblast_nt_vector_clas_>5.
 ↪png')
```



```
[61]: ax = df['class'].value_counts().loc[lambda x : x<=5].plot(kind='bar',
                                                    figsize=(20,6),
                                                    title="Vector contig counts")
      ax.set_xlabel("Vector class")
      ax.set_ylabel("Frequency")
      plt.tight_layout()
      plt.
       ↪savefig(path+'PRJNA607174_contig_BLAST_analysis_vector_magicblast_nt_vector_clas_<5.
       ↪png')
```



```
[ ]:
```